

AP Statistics Chapter 1 Notes - Exploring Data

1.1/2: Categorical Variables and Displaying Distributions with Graphs

Individuals and Variables

- **Individuals** are objects described by a set of data. Individuals may be people, but they may also be animals or things.
- A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Categorical and Quantitative Variables

- A **categorical variable** places an individual into one of several groups or categories.
- A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

Distribution

The **distribution** of a variable tells us what values the variable takes and how often it takes these variables.

Describing the Overall Pattern of a Distribution – Remember your SOCS

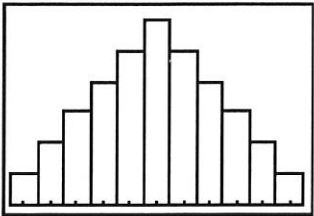
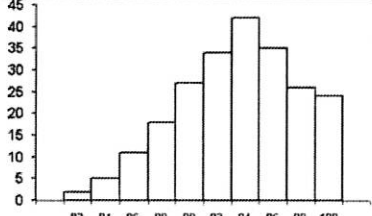
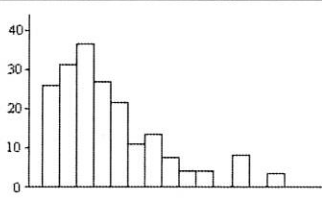
To describe the overall pattern of a distribution, address all of the following:

- **Spread** – give the lowest and highest value in the data set
- **Outliers** – are there any values that stand out as unusual?
- **Center** – what is the approximate average value of the data (only an estimation)
- **Shape** – does the graph show symmetry, or is it skewed in one direction (see below)

Outliers

An outlier in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Describing the SHAPE of a distribution – Symmetric and Skewed Distributions

Symmetric	Skewed Left	Skewed Right
		
Mean = Median	Mean < Median	Mean > Median

Time Plot

- A **time plot** of a variable plots each observation against the time at which it was measured.
- Always mark the time scale on the horizontal axis and the variable of interest on the vertical axis. If there are not too many points, connecting the points by lines helps show the pattern of changes over time.

1.3: Describing Distributions with Numbers

The Mean (\bar{x})

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are x_1, x_2, \dots, x_n , their mean is:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad \text{or simply,} \quad \bar{x} = \sum_{i=1}^n x_i$$

The Median (M)

- The median M is the midpoint of distribution, the number such that half the observations are smaller and the other half are larger. To find the median of distribution:
- Arrange all observation in order of size, from smallest to largest.
- If the number of observations n is odd, the median M is the center observation in the ordered list. The position of the center observation can be found at $(n + 1) / 2$
- If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The position of the two middle values are $n/2$ and $n/2 + 1$

The Five-Number Summary

The five-number summary of a data set consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is:

Minimum – Q_1 – M – Q_3 – Maximum

The Quartiles (Q_1 and Q_3)

- To calculate the quartiles, arrange the observations in increasing order and locate the median M in the ordered list of observations.
- The 1st quartile (Q_1) is middle number of the values that are less than the median.
- The 3rd quartile (Q_3) is the middle number of the values that are greater than the median.

Example

2	14	28	29	30	32	33	34	40	42	52
Min		Q1			Med			Q3		Max

The Interquartile Range (IQR)

The IQR is the distance between the first and third quartiles, $IQR = Q_3 - Q_1$

Outliers: The 1.5 x IQR Criterion

Call an observation an outlier if it falls more than $1.5 \times IQR$ below the first quartile or above the third quartile. Using the 5-number summary from above as an example ($IQR = 40 - 28 = 12$)

- Low outlier cutoff: $Q_1 - 1.5 \times IQR$ (example: $28 - 1.5(12) = 28 - 18 = 10$) Therefore, the 2 is an outlier.
- High outlier cutoff: $Q_3 + 1.5 \times IQR$ (example: $40 + 1.5(12) = 40 + 18 = 58$) no outlier

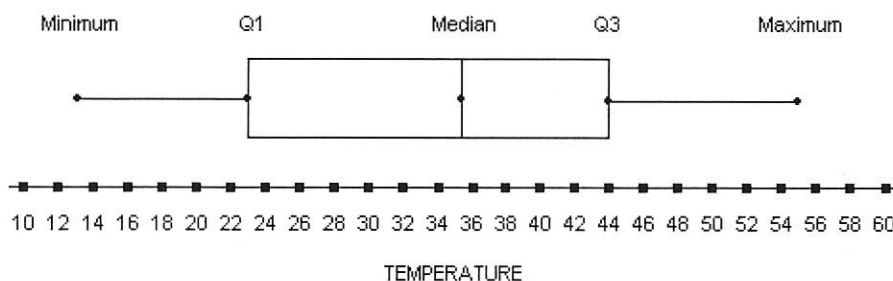
1.3: Describing Distributions with Numbers

Boxplot

A boxplot is a graph of the five-number summary, with outliers plotted individually.

- A central box spans the quartiles.
- A line in the box marks the median.
- Observations more than 1.5 x IQR outside the central box are plotted individually.
- Lines extend from the box out to the smallest and largest observations, not the outliers.

Example:



The Standard Deviation (S or Sx)

The standard deviation of a set of observations is the average of the squares of the deviations of the observations from their mean. The formula for the standard deviation of n observations x_1, x_2, \dots, x_n is:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Calculation of the Standard Deviation

Consider the data below which has a mean of 4.8:

x_i	$x_i - \text{mean}$	$(x_i - \text{mean})^2$
6	$6 - 4.8 = 1.2$	$(1.2)^2 = 1.44$
3	$3 - 4.8 = -1.8$	$(-1.8)^2 = 3.24$
8	$8 - 4.8 = 3.2$	$(3.2)^2 = 10.24$
5	$5 - 4.8 = 0.2$	$(0.2)^2 = 0.04$
2	$2 - 4.8 = -2.8$	$(-2.8)^2 = 7.84$
Sum	0	22.8

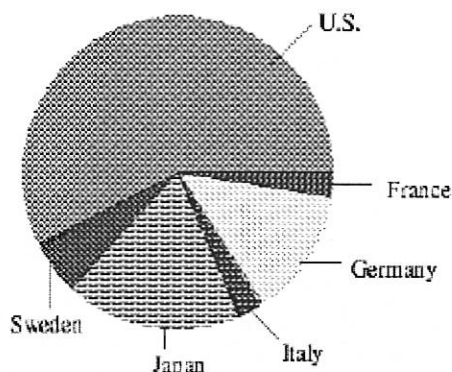
So the standard deviation is $\sqrt{22.8 / (5 - 1)} = \sqrt{22.8 / 4} = \sqrt{5.7} = 2.387$

AP Statistics Chapter 1 Test Study Guide

Multiple Choice

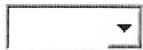
Identify the choice that best completes the statement or answers the question.

- ☐ 1. You measure the age, marital status and earned income of an SRS of 1463 women. The number and type of variables you have measured is
- 1463; all quantitative.
 - four; two categorical and two quantitative.
 - four; one categorical and three quantitative.
 - three; two categorical and one quantitative.
 - three; one categorical and two quantitative.
- ☐ 2. Consumers' Union measured the gas mileage in miles per gallon of 38 1978–1979 model automobiles on a special test track. The pie chart below provides information about the country of manufacture of the model cars used by Consumers Union. Based on the pie chart, we may conclude that:

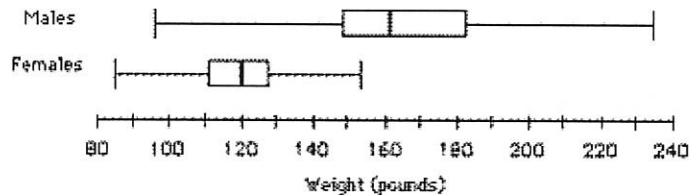


- Japanese cars get significantly lower gas mileage than cars of other countries. This is because their slice of the pie is at the bottom of the chart.
 - U.S. cars get significantly higher gas mileage than cars from other countries.
 - Swedish cars get gas mileages that are between those of Japanese and U.S. cars.
 - Mercedes, Audi, Porsche, and BMW represent approximately a quarter of the cars tested.
 - More than half of the cars in the study were from the United States.
- ☐ 3. "Normal" body temperature varies by time of day. A series of readings was taken of the body temperature of a subject. The mean reading was found to be 36.5°C with a standard deviation of 0.3°C (recall that $^{\circ}\text{F} = ^{\circ}\text{C}(1.8) + 32$). When converted to $^{\circ}\text{F}$, the mean and standard deviation are:
- 97.7, 32
 - 97.7, 0.30
 - 97.7, 0.54
 - 97.7, 0.97
 - 97.7, 1.80
- ☐ 4. Which of the following is likely to have a mean that is smaller than the median?
- The salaries of all National Football League players.

- b. The scores of students (out of 100 points) on a very easy exam in which most get nearly perfect scores but a few do very poorly.
- c. The prices of homes in a large city.
- d. The scores of students (out of 100 points) on a very difficult exam in which most get poor scores but a few do very well.
- e. Amounts awarded by civil court juries.

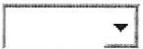


5. The weights of the male and female students in a class are summarized in the following boxplots:

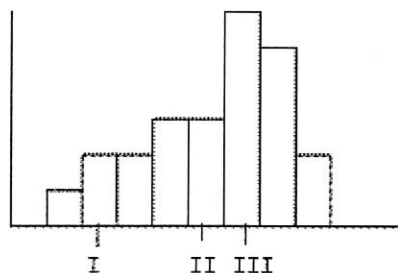


Which of the following is NOT correct?

- a. About 50% of the male students have weights between 150 and 185 pounds.
- b. About 25% of female students have weights more than 130 pounds.
- c. The median weight of male students is about 162 pounds.
- d. The mean weight of female students is about 120 pounds because of symmetry.
- e. The male students have less variability than the female students.



6. For the following histogram, what is the proper ordering of the mean, median, and mode? Note that the graph is NOT numerically precise—only the relative positions are important.



- a. I = mean, II = median, III = mode
- b. I = mode, II = median, III = mean
- c. I = median, II = mean, III = mode
- d. I = mode, II = mean, III = median
- e. I = mean, II = mode, III = median



7. A medical researcher collects health data on many women in each of several countries. One of the variables measured for each woman in the study is her weight in pounds. The following list gives the five-number summary for the weights of women in one of the countries.

Country A: 100, 110, 120, 160, 200

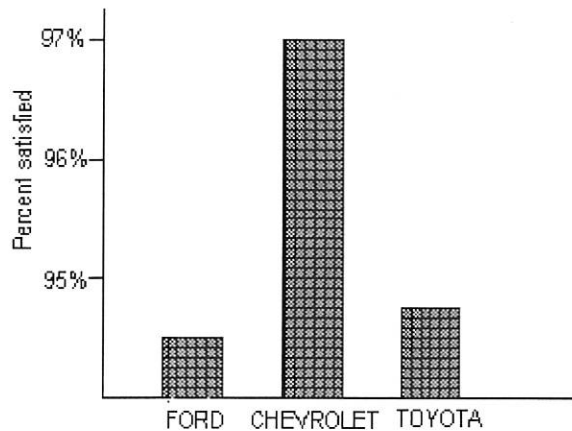
About what percentage of Country A women weigh between 110 and 200 pounds?

- a. 50%
- b. 65%
- c. 75%

- d. 85%
- e. 95%

☐

8. The following bar graph gives the percent of owners of three brands of trucks who are satisfied with their truck.



From this graph we may legitimately conclude that:

- a. Owners of other brands of trucks are less satisfied than the owners of these three brands.
- b. Chevrolet owners are substantially more satisfied than Ford or Toyota owners.
- c. There is very little difference in the satisfaction of owners for the three brands.
- d. Chevrolet probably sells more trucks than Ford or Toyota.
- e. A pie chart would have been a better choice for displaying this data.

☐

9. A sample of 99 distances has a mean of 24 feet and a median of 24.5 feet. Unfortunately, it has just been discovered that an observation which was erroneously recorded as "30" actually had a value of "35". If we make this correction to the data, then:

- a. The mean remains the same, but the median is increased.
- b. The mean and median remain the same.
- c. The median remains the same, but the mean is increased.
- d. The mean and median are both increased.
- e. We do not know how the mean and median are affected without further calculations, but the variance is increased.

☐

10. The five-number summary for scores on a statistics exam is 11, 35, 61, 70, 79. In all, 380 students took the test. About how many had scores between 35 and 61?

- a. 26
- b. 76
- c. 95
- d. 190
- e. None of these

AP Statistics – Chapter 1 Free Response Practice Test

1. The test grades for a certain class were entered into a Minitab worksheet, and then “Descriptive Statistics” were requested. The results were

MTB > Describe 'Grades'.

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
Grades	28	74.71	76.00	75.50	12.61	2.38
	MIN	MAX	Q1	Q3		
Grades	35.00	94.00	68.00	84.00		

- (a) Determine the IQR for this data.
- (b) Using the answer from part (a), determine whether the lowest and highest values in the data are outliers.
2. The following data represent scores of 50 students on a calculus test.

72	72	93	70	59	78	74	65	73	80
57	67	72	57	83	76	74	56	68	67
74	76	79	72	61	72	73	76	67	49
71	53	67	65	99	83	69	61	72	68
65	51	75	68	75	66	77	61	64	74

- (a) Construct a *frequency* histogram for this data set.
- (b) Describe the shape, center, and spread of the distribution of test scores.
3. During the early part of the 1994 baseball season, many sports fans and baseball players noticed that the number of home runs being hit seemed to be unusually large. Here are the data on the number of home runs hit by American and National League teams:

American League	35, 40, 43, 49, 51, 54, 57, 58, 58, 64, 68, 68, 75, 77
National League	29, 31, 42, 46, 47, 48, 48, 53, 55, 55, 55, 63, 63, 67

- (a) Construct a back-to-back stemplot to compare the number of home runs hit in the two leagues.
- (b) Write a few sentences comparing the distributions of home runs in the two leagues. Be sure to include a comparison of the medians as part of your discussion.